



Summary

Contents

A Quarterly Information Bulletin from the Methodology Division

Contents

ABSEst: New Estimation Facilities for ABS Surveys
Research into Venture Capital
Remote Access to ABS Microdata
Confidentialising Microdata : Aims
Standard Level of Detail for Confidentialised Unit Record Files
Phase-in of Computer Assisted Interviewing to the Labour Force Survey
Innovation Survey Design
Maintaining Optimal Frame Coverage using Administrative Data

ABSEst: New Estimation Facilities for ABS Surveys

ABSEst: New Estimation Facilities for ABS Surveys

The introduction of a New Tax System in Australia in July 2000 has enabled the ABS to access Business Activity Statement (BAS) data for most Australian businesses. This data can be used as auxiliary information to reduce cost, provider load and/or sampling errors for many ABS business surveys. A new estimation facility, ABSEst, has been developed to enable maximum use of the auxiliary data through alternative estimation methodologies.

ABSEst is a collection of linked software components. The most important is a generalized regression estimation (GREG) component. This component applies the GREG methodology, which enables the flexible use of auxiliary information using general linear models. If the auxiliary information is correlated with survey variables, then the GREG methodology will improve the accuracy of the estimates for the current sample sizes or reduce the current sample sizes with no reduction in the accuracy of the estimates.

Other software components include: automated outlier detection and treatment using winsorization; and variance estimation using a half-sample bootstrap method.

A number of methodological innovations were made as part of the ABSEst development. The existing winsorization techniques needed to be extended to deal with linear models, using outlier-robust model fitting. A "zeroes-adjusted" GREG estimator was developed to provide more accurate estimates for surveys with many zero values due to defunct and out of scope businesses.

The ABSEst components will be introduced to the Monthly Retail Trade Survey in the March quarter 2004. The introduction of the GREG methodology, as well as redesigning the sample, is expected to result in a 16% reduction in sample size while also slightly improving the precision of the estimates.

ABSEst is expected to be applied to other ABS business surveys over the next few years, as well as be an alternative to GREGWT currently used for household surveys.

If you have any queries on the ABSEst development, please contact: John Preston on (02) 6252 6970.

Email: john.preston@abs.gov.au

Research into Venture Capital

Research into Venture Capital

The Venture Capital (VC) market is an emerging investment sector in Australia. Venture capital is defined as capital invested in high risk enterprises that promise high returns. While various government programs have been in effect since the early 1980's to stimulate the venture capital sector in Australia, very little is known about the dynamics of the VC market due to a lack of relevant statistics. Even though various organisations gather and publish VC data, the ABS is the first statistical agency to conduct a comprehensive survey. The ABS survey collects most of the activities involved in the VC process including information on investors, funds, fund managers and investees.

The ABS began its first VC survey for the 1999-2000 financial year at the request of the Department of Industry, Tourism and Resources (DITR) and the National Office for the Information Economy (NOIE). Since then the ABS has conducted 3 annual surveys. In the most recent survey conducted for the financial year ending June 2002, VC funds attracted \$6.9 billion in capital commitments (capital promised) and invested \$4.4 billion. The availability of comprehensive VC survey data generated considerable interest from a wide range of users including our major clients (DITR, NOIE) and others (Reserve Bank of Australia, Treasury, ACT government, academics etc.) for policy purposes and research.

Recently, Analytical Services Branch (ASB) conducted two exploratory studies using VC survey data to widen the knowledge of VC needs and to add value to survey statistics.

Identify determinants of VC fundraising and investment performance

- The objective of this exploratory project was to determine statistically significant relationships in VC investor behaviour (factors influencing commitments), funds behaviour (factors influencing drawdowns) and investee behaviour (deals performance). Empirical models were developed to represent the expected relationships. These were then tested using the third-wave of VC survey data. The results showed that there were factors that are associated with VC fundraising and investment performance. A paper summarising the findings of this research was presented at the 16th Australasian Finance and Banking Conference in Sydney.

Develop a performance indicator

- A range of methodologies were investigated to develop a performance indicator for VC investments. The development of a performance indicator provides a statistical measure of performance of VC enterprises and thus assists investors with their decision making. Three methodologies were tested: the application of portfolio allocation method; discounting method; and weighted standard deviation method. The results showed that different industries and activities in VC have different levels of returns and risks. The portfolio allocation method provided a better measure of performance than other methods.

Further research that could be undertaken using the VC data includes linking VC data with other business survey data, longitudinal analysis, VC exit pattern and alternative methods for analysing investment performance.

For more information, please contact Tala Talgaswatta on 02 6252 5376 or Luke Samy on 02 6252 5933.

Email: tala.talgaswatta@abs.gov.au or
luke.samy@abs.gov.au.

Remote Access to ABS Microdata

Remote Access to ABS Microdata

The ABS has recently expanded its ability to provide microdata to external researchers. Access to confidentialised unit record files (CURFs) has historically been limited to basic information disseminated via CD-ROM. In response to demand for more detailed microdata, the ABS is now providing access to CURFs with expanded data via the Remote Access Data Laboratory (RADL). These expanded CURFs are released only via the RADL, and contain extra data items and more detail than basic CURFs. Basic CURFs are available on CD-ROM and may also be accessed via the RADL.

The RADL was launched in November 2003. It is a web-based statistical tool which enables academics, researchers and other government agencies to carry out batch-mode statistical analysis from their desktops in remote locations. Access to the RADL is via the ABS website, with all RADL microdata, infrastructure and associated systems situated behind the ABS firewall.

Coupled with the ABS' desire to release CURFs is a requirement to protect the privacy and confidentiality of our data providers. Use of the RADL is therefore restricted to statistical use of the microdata and is subject to the following confidentiality protections and regulations.

- **Undertakings:** RADL clients and their organisations must sign an undertaking. The undertaking is a legal document where the client describes their proposed analysis and agrees to comply with a set of microdata security principles.
- **Education:** RADL Clients must read the Responsible Access to ABS Confidentialised Unit Record Files Training Manual to ensure that they understand their microdata security responsibilities.
- **Confidentialised Data:** The microdata which are available on the RADL have been confidentialised to protect against the risk of spontaneous recognition. The confidentialisation process includes the removal of name and address information as well as any or all of: reduction of available detail; data swapping; perturbation; or removal of very unusual records.
- **Automatic Checks:** The RADL automatically checks for large scale printing of unit record data, as well as the size of any output tables produced. In the future there will be differential print limits as well as sparsity checks on all output tables.
- **Audits:** All RADL jobs and their associated files are logged and retained for auditing purposes. All programs are examined for compliance with RADL guidelines, and results are checked for identifying information. Clients are notified retrospectively if their programs have breached the guidelines, or if their results file must be kept secure, or destroyed.

As of November 2003 there were 31 organisations and 144 clients with RADL access. CURF access has steadily grown, with the National Health Survey 2001 Expanded CURFs being used most extensively. An expanded 1997 Time Use Survey CURF, the expanded 2001 Population Census Household Sample File, and the expanded 2002 National Health Survey (Indigenous) CURF are currently available via the RADL. The 1999-2000 Income and Housing Costs Survey Basic CURF is also available via the RADL and also on CD-ROM. Other CURFs to be released via the RADL over the next six months include basic and expanded CURFs from the 2002 General Social Survey, and the CURF from the 2002 Crime and Safety Survey. For information on these new files refer to the Access to CURFs pages on the ABS web site (www.abs.gov.au).

For more information, please contact: Dale Wallace on (02) 6252 7313.

Email: d.wallace@abs.gov.au

Confidentialising Microdata : Aims

Confidentialising Microdata : Aims

To confidentialise microdata, the ABS aims to protect the data against two specific disclosure risk scenarios: spontaneous recognition and matching to lists.

Spontaneous recognition occurs when a user, while looking at information on a microdata file, happens to notice some unusual characteristics which remind her/him of someone that they know. Looking further at other data items can confirm the identity.

List matching occurs when two datasets could be matched via common variables and, in cases where the matches are one-to-one, the combined set of variables pose an additional disclosure risk. As administrative lists often contain similar demographic variables (age, sex, geographic location, etc.) as the ABS microdata, the risk of list matching to administrative lists is assessed.

For basic CURFs, the aim is to protect against both the spontaneous recognition and the list matching risk. For expanded CURFs, the aim is to protect the data against spontaneous recognition only, as extra protections are built into the Remote Access Data Lab infrastructure to protect list matching. The 2004 work program for the Data Access and Confidentiality Methodology Unit includes conducting research to improve methods for assessing and addressing the spontaneous recognition and list matching risks, as well as developing software tools to make these processes more efficient.

For more information about confidentiality, please contact Kirsty Leslie on (02) 6252 5594 or Paul Schubert (02) 6252 7306.

Email: kirsty.leslie@abs.gov.au or
paul.schubert@abs.gov.au

Standard Level of Detail for Confidentialised Unit Record Files

Standard Level of Detail for Confidentialised Unit Record Files

The ABS makes unidentifiable microdata from its surveys available to users in the form of Confidentialised Unit Record Files (CURFs). In the past, the balance between maximising the level of detail and maintaining confidentiality has been done separately for each individual CURF. This has resulted in trade-offs in the level of detail between variables to best meet the needs of clients for a particular CURF. This method has been labour intensive, and while it has given some flexibility for client areas, it has resulted in inconsistency between collections and over time, leading to a lack of predictability for what will be allowable on the CURF. It was therefore decided to standardise the level of detail for CURFs in order to reduce the effort required in producing and assessing CURFs.

The nature of different collections and the different release mechanisms has dictated that a number of different standard levels of details be developed depending on the type of collection, and whether the CURF is to be released on CD ROM and the remote access data lab (RADL), or just on the RADL.

Basic files (CD ROM or RADL release)

- income, labour and education CURFs to contain more detail for age around the 'transition' age groups, (i.e. 15-24 and 55-64 years):
- other social surveys may have children in scope so a different standard level of detail was needed. It was noted that in the past some collections had opted for different Geography to be included on CURFs for different reasons. As a result, two 'options' for Geography are considered for this standard - state and one sub-state geography, and remoteness and socioeconomic indexes for areas (SEIFA) presented in quintiles:
- the Population Census 1% Household Sample File has a sample of at least four times that of most other collections, thus supporting more detailed analysis. A separate standard level of detail was considered for the Census file.

Expanded files (RADL release only)

- for most expanded CURFs, the level of detail of variables is that which would be useful for analysis. Variables which identify groups of people that a user is likely to know, (e.g. geographic area or industry), are not as fine as other variables. In addition, masking of individual records (through altering values of variables for a small number of particularly unusual records) on the file is the main strategy for protection against spontaneous recognition;
- again, the Census file supports more detailed analysis, so a separate standard level of detail was required;
- Indigenous collections were also considered as a separate case, as the size and distribution of the Indigenous population would mean that some variables would need to be further restricted from that proposed in the expanded standard.

So far, efforts to standardise have been concentrated on a small set of core person level variables which have, in the past, most commonly had the level of detail collapsed to reduce the risk of identification. The standard variables can be grouped according to whether they pose a risk to list matching, to spontaneous recognition, or to both.

Risk of list matching and spontaneous recognition:

- Sex;
- Age;
- Marital Status;
- Country of Birth;
- Year of arrival;
- Geography;
- Indigenous status.

Risk of list matching only:

- SEIFA.

Risk of spontaneous recognition only:

- Education;

- Industry;
- Occupation.

Income has also been considered as a risk for list matching and spontaneous recognition, but due to some complications in determining the appropriate level of detail on non-income survey CURFs, it has not been included in the first set of variables to be signed off.

Approval and future developments

Only standards for household collections have been developed so far, as almost all CURFs produced by the ABS have been from household collections. It is planned to develop standard levels of details for other common person and household level variables for CURFs in the near future. Plans to develop standards for business surveys as experience with them expands are also being developed.

All CURFs will be subject to the appropriate standards unless there is very well justified user demand and an appropriate trade off in the level of detail is made.

An assessment of the combined disclosure risk posed by the variables in the first set of standards has been conducted and considered by the Micro Data Review Panel. Approval for the standards will be sought from the Australian Statistician.

For more information, including the complete detail for the variables in the standards, please contact Kirsty Leslie on (02) 6252 5594 or Paul Schubert (02) 6252 7306.

email: kirsty.leslie@abs.gov.au or
paul.schubert@abs.gov.au

Phase-in of Computer Assisted Interviewing to the Labour Force Survey

Phase-in of Computer Assisted Interviewing to the Labour Force Survey

Prior to 2003, Labour Force Survey (LFS) interviewers have made use of paper forms to collect information from respondents. The ABS is in the process of phasing-in the use of Computer Assisted Interviewing (CAI) for the LFS between October 2003 and June 2004. In October 2003, 10% of interviewers made use of computers for the first time (i.e. about 6,000 persons or interviews). The remaining interviewers will be phased-in from paper forms to CAI over 2004.

A delay between the first and second phase-in groups was designed to minimise risks associated with the computing systems: if there were system problems associated with the 10% sample in October, then there would be 3 months to fix them before the next, much larger, phase-in of interviewers scheduled for February. Also, the first phase-in group was small in size in order to mitigate the risk of a substantial mode effect. In such a situation, dropping the CAI sample (10%) of respondents would only have a minor affect on the coherence of the LFS time series. Lastly, the period of the phase-in was limited in order to minimise the period of uncertainty affecting the LFS series.

Given the small size of the first phase-in group, it was accepted that a statistical test for a mode effect would have poor accuracy and low power.

Household Survey Methodology's (HSM) involvement in this project has been to:

- randomly select interviewers to one of the phase in groups;
- maintain a frame of interviewers in order to make these selections;
- ensure that the selected interviewers in each group are allocated a representative sample of workloads, in terms of location and the LF status of the persons interviewed;
- measure the size of the mode effect at various levels (Australia, state, part of state), by characteristics of the respondents (age and sex), characteristics of the interviewer (experience of interviewer) and characteristics of the interview (whether the interview took place face-to-face or over the telephone);
- isolate potential sources of the mode affect, in order to inform users, and to determine if the mode effect is a phantom affect (i.e. due to non-CAI sources).

Composite estimation is being used to assess the mode effect. A composite estimator makes use of more than one month of LFS data. In this case, to estimate a potential mode effect in the October LFS estimates, the composite

estimator makes use of August, September and October LFS data.

Investigations into the existence of a possible mode effect are continuing. Four months of data have been analysed so far, but the results of these investigations are less than definitive given the size of the sample. Further analysis following the phase-in of the second group is expected to be more conclusive.

For more information, please contact James Chipperfield (02) 6252 7301.

Email: james.chipperfield@abs.gov.au.

Innovation Survey Design

Innovation Survey Design

In response to strong user demand for updated and internationally comparable statistics covering technological and non-technological innovation for the goods producing and services sector, the ABS is conducting an Innovation Survey in 2004. The scope of the survey is all businesses with employment of 5 or more on the business register in all ANZSIC divisions with the exception of Agriculture, Government, Health & Community Personal & Other.

The survey was designed to meet accuracy constraints on two main outputs: the expenditure on innovation; and the proportion of innovative businesses. Information on the expenditure on innovation was available at the unit record level for the Manufacturing ANZSIC only, taken from the 1996/97 ABS Innovation Survey. Information on the proportion of innovative businesses was available at the ANZSIC by size level (a unit record file was not available), from the 1993/94 Innovation in Industry survey.

The design constraints were a 10% RSE for the expenditure on innovation and a 0.05 standard error for the proportion of innovators at the ANZSIC level (2-digit ANZSIC level within Manufacturing and Property and Business Services). Estimated stratum level population variances were derived for expenditure on innovation in the Manufacturing ANZSIC using the unit record file from the 1996/97 Innovation survey. Stratum level proportions were set to that at the ANZSIC by size level published from the 1993/94 Innovation in Industry survey, for all strata within that ANZSIC by size class (i.e. there was no difference across states). Population counts were taken from an appropriately scoped ABS business surveys frame used for the 2002/03 annuals.

Deriving separate allocations to meet these two constraints in the Manufacturing ANZSIC resulted in quite different sample sizes, with the sample needed to meet the expenditure RSEs being much larger than that needed to meet the proportion standard errors. This suggested that using the allocation derived for the proportion of innovators in the non-Manufacturing industries would not result in reliable expenditure data. However, since there was no design information for expenditure available in the non-Manufacturing industries the only option was to inflate the allocation derived for the proportion of innovators. To do this, a relationship between the allocation for expenditure and proportion within the Manufacturing ANZSIC was determined. This was then applied to the allocation for the proportion of innovators in the non-Manufacturing ANZSICs. The total allocation then comprised this inflated allocation in the non-Manufacturing ANZSICs plus the allocation for expenditure in Manufacturing.

This final allocation was below the affordable level determined by the Business Statistics Centre (BSC), so given the age and the restricted nature of the design data it was decided to further inflate the allocation to this affordable level. This was done by allocating extra sample to those state by industry levels where the expected accuracy was the lowest.

It has been stressed to the BSC, National Statistics Centre (NSC) and external advisory group that due to the lack of design data plus the assumptions that were made in the design process, the design RSEs and standard errors are indicative only and can not be guaranteed.

The survey forms are due to be despatched in February 2004, with a publication date planned for November 2004.

For more information please contact Helen Teasdale on 08 9360-5991.

Email : helen.teasdale@abs.gov.au.

Maintaining Optimal Frame Coverage using Administrative Data

Maintaining Optimal Frame Coverage using Administrative Data

The ABS Business Register is essentially a list of all employing and non-employing businesses operating in

Australia. The Business Register provides the statistical framework from which representative samples are drawn for most ABS economic censuses and sample surveys. Therefore, it is important that the Business Register is a comprehensive, accurate and regularly updated source of businesses.

However, financial constraints as well as practical limits on the source data available for updating the Business Register mean that the Business Register can never be perfect and completely up-to-date. The ABS has strategies in place to deal with imperfections on the Business Register. These strategies must be operationally practical, as well as theoretically sound.

Errors caused by imperfect registers fall into two types:

- **content** errors refer to inaccuracies in the information recorded about units on the register. Usually their effect can readily be estimated from the sample, provided that there are procedures in place which can consistently and reliably identify these register imperfections in the sample;
- **coverage** errors relate to the presence or absence of units on the register. Typically they contribute mostly to the fixed error or bias in the estimate and usually additional investigations are needed to estimate their effect. New business provisions are used to deal with undercoverage for businesses not yet added to the register. A technique to reduce overcoverage on an ongoing basis is to remove businesses from the frame that are defined as 'long term non-remitters'.

Having businesses that are no longer operating or businesses that are out of scope of our surveys on our frames reduces sampling efficiency. The ABS uses administrative taxation data to identify and remove from frames such businesses. The main criterion in determining long term non-remitters is that businesses have not remitted taxation data for the last five consecutive quarters. One reason why five consecutive quarters was chosen was because some businesses only need to provide the Australian Taxation Office with taxation data on an annual basis.

As the ABS uses a technique that reduces overcoverage, there is actually a chance that in scope operating businesses are removed in error. The main reason for this error is that some businesses are late in providing taxation data to the Australian Taxation Office.

An investigation conducted by the Victorian Methodology Unit looked into whether having a five quarter non-remittance period was sufficient to determine long term non-remitters. It was concluded that extending the non-remittance period beyond five quarters would result in fewer businesses being incorrectly identified as non-remitters. However, it would also lead to a reduction in the number of 'true' long term non-remitters identified, resulting in increased sample error. As the increase in sample error outweighed the benefits from reducing the error due to undercoverage, a decision was made to stay with the five quarter non-remittance period.

For more information, please contact: Rosslyn Starick on 03 9615 7689.

Email: rosslyn.starick@abs.gov.au

About this release

Methodological News

© Commonwealth of Australia

All data and other material produced by the Australian Bureau of Statistics (ABS) constitutes Commonwealth copyright administered by the ABS. The ABS reserves the right to set out the terms and conditions for the use of such material. Unless otherwise noted, all material on this website – except the ABS logo, the Commonwealth Coat of Arms, and any material protected by a trade mark – is licensed under a Creative Commons Attribution 2.5 Australia licence